Submission in Response to NSF CI 2030 Request for Information

DATE AND TIME: 2017-04-06 02:30:31
REFERENCE NO: 310

Author Names & Affiliations

- Eli Dart - ESnet, Lawrence Berkeley National Laboratory

Contact Email Address (for NSF use only)

(Hidden)

Research Domain, discipline, and sub-discipline

Computer Science, Network Engineering, Cyberinfrastructure

Title of Submission

Cyberinfrastructure for data-intensive science

Abstract (maximum ~200 words).

Many different fields of science and engineering are facing a common set of challenges involving data analysis, data management, data movement, and data storage and archiving. To address these needs for all fields, a whole-ecosystem approach to solving these data challenges is required. The successful approach will consist of several synergistic and complementary initiatives, including: building on the success of the CC* and DIBBs programs, using the cyberinfrastructure they deployed as a platform for higher-level services; integrating HPC capabilities to the data ecosystem, including campus, regional and national capabilities; making computing and data capabilities easier to use so that the burden of integration does not fall on the scientist; developing and sustaining the professions which span the traditional gap between science and cyberinfrastructure (variously called Cyberinfrastructure Practitioners, Cyberinfrastructure Engineers, Research Facilitators, Science Engagement Engineers, and the like); and continuing efforts to diversify the scientific workforce so that the best minds in the country can find a home in the sciences and so give their best to their fields. This is a complex task, but its successful completion will transform the productivity of the US science complex, and accelerate the discoveries needed to implement the next-generation solutions to today's challenges.

Question 1 Research Challenge(s) (maximum ~1200 words): Describe current or emerging science or engineering research challenge(s), providing context in terms of recent research activities and standing questions in the field.

Rather than describe specific research challenges in a particular field, we instead take a broad view with the intent of identifying a set of similar challenges facing many fields. In many areas of science, the productivity of a scientist or science collaboration is closely tied to the ability to gain insight from scientific data. Unfortunately, despite the promise of the increasing resolution, diversity, and scale of scientific data available to researchers, the mundane mechanics of transferring, translating, transforming, and managing data make it very difficult for many scientists to take full advantage of the data available. Indeed, scientists will use instruments at less than full capability simply because they don’t have the resources to manage the scale of data produced by a new instrument at full resolution. Broadly-applicable solutions to these problems would serve a great many scientists in a large number of fields.

Specifically, a critical need exists for data transfer, data sharing, and data storage. Data analysis doesn’t work very well if the analyst can’t get the data to the computing system that will be used for the analysis, and if the results cannot be shared then collaboration becomes difficult. If the data sets and the results of the analysis cannot be stored, subsequent analysis and reproducibility become difficult or impossible. These needs can be addressed by a relatively small number of cyberinfrastructure components and services, and their broad deployment will help a large number of scientists in a wide variety of fields.

Question 2 Cyberinfrastructure Needed to Address the Research Challenge(s) (maximum ~1200 words): Describe any limitations or absence of existing cyberinfrastructure, and/or specific technical advancements in cyberinfrastructure (e.g. advanced computing, data infrastructure, software infrastructure, applications, networking, cybersecurity), that must be addressed to accomplish the identified research challenge(s).

Cyberinfrastructure has a critical role to play in addressing the needs of the data space. Investments by NSF have provided a solid foundation upon which to build, but more remains to be done. In particular, the Science DMZs which have been deployed at universities and the high-speed networks which interconnect them together form a platform upon which high-performance, scalable data services and capabilities can be built.

The gap between existing cyberinfrastructure capabilities and the broad needs of many science collaborations can be expressed in terms of a common workflow design pattern. Imagine a scientist who makes use of a scientific instrument to collect data which describe a set of samples. These could be bone fragments, protein crystals, agricultural samples, or chemical processes. The scientists collects a large volume (many terabytes) of high-resolution data. In order to effectively analyze the data, a HPC system is required - such analysis is beyond the capabilities of the scientist’s laptop. The scientist must then transfer the data to a HPC facility, possibly at the laboratory which houses the instrument but increasingly at a campus or national facility. Once the data sets are analyzed, the results can be shared with colleagues, published, analyzed later in combination with other data, or whatever is appropriate. However, that requires that the data sets be stored in a location where they can be accessed later - later accessibility includes not just the ability to download the data from the storage system but the ability to search for and find the data set again.

This template example illustrates several points. First, the basic components of the example - instrument/experiment, data transfer/movement, data analysis, data sharing, and data storage/archiving - are part of a wide variety of disciplines. Biology, materials science, chemistry, genomics, and physics are a few examples of the fields which share this basic abstract workflow design pattern. Second, in today’s environment, the scientist is responsible for integrating all these various components into a workflow that is productive. As each component of the workflow has become more sophisticated and capable, it has also become more complex - the only way for a scientist (who is an expert in his or her area of specialization and typically not in cyberinfrastructure engineering) to use the cyberinfrastructure components effectively is for those cyberinfrastructure components to have standardized interfaces that are straightforward to use without the need to understand how the cyberinfrastructure capability is implemented. Third, this example holds for a wide variety of use cases. The data might come from a gene sequencer, data repository, telescope, electron microscope, or HPC simulation, and the particular data analysis code will certainly vary from collaboration to collaboration. However, when a person needs to analyze the data, this common design pattern matches the workflow of a great many scientists.

This, then, is the task: scientists must have routinely available to them an operational set of capabilities for acquiring, transferring, analyzing, managing, storing, sharing, and archiving data at a scale which matches their needs and keeps pace with the state of the art. Several cyberinfrastructure programs have been very successful - examples include CC* and DIBBs - and these programs have built the foundation for the next layer of capabilities. The next steps involve integrating HPC analysis capabilities with experiments - not just via Science DMZs connected to networks, but via interoperable software which can effectively tie the experiment, the Science DMZs, the networks, and the HPC center together into a coherent whole. The scientist must be able to use this super-capability without being an expert in the implementation of its components - there just aren’t enough hours in the day to look “under the hood” of all the pieces or troubleshoot their behavior. The parts of the whole have to routinely function properly in the general case so that the scientist can use the cyberinfrastructure as a productive capability which provides a platform for the next generation of discoveries.

Question 3 Other considerations (maximum ~1200 words, optional): Any other relevant aspects, such as organization, process, learning and workforce development, access, and sustainability, that need to be addressed; or any other issues that NSF should consider.

Building and operating the cyberinfrastructure to provide a the underlying components of a successful data analysis ecosystem carries with it several challenges.

The workforce which will build and operate these capabilities is a key component of their success, and there are some important gaps to be filled. Capable cyberinfrastructure is complex and interdomain (by this we mean that systems which must work seamlessly together are built from components and services which are run by multiple collaborating organizations and interconnected by science networks). There exists a little-known profession which keeps cyberinfrastructure capabilities interoperating well and functioning together as a coherent whole. These professionals are called Cyberinfrastructure Engineers, Research Facilitators, Science Engagement Engineers, and the like. Their skillset includes network engineering, network and systems security, programming, data analysis, and (critically) verbal and written interpersonal communication. These are the people who ensure that the larger system functions well as a whole, and these are the people who help scientists solve integration problems. This profession needs to be codified and formalized so that the people who work in it have a career path which makes investing in the necessary multidisciplinary expertise worthwhile.

In addition to formalizing the profession of CI engineering, workforce diversity efforts must continue. STEM professions are terribly unbalanced in terms of gender, race, and many other important diversity metrics. This means that not only are valuable viewpoints underrepresented in the profession, but the profession is missing the benefit of bright minds which would otherwise be delighted to find a home there. It should also be pointed out that race and gender imbalance is a matter of justice - people who would find joy and fulfilment by devoting their professional lives to serving the quest for scientific knowledge are being excluded because of who they are. This is wrong on human grounds, regardless of the additional deleterious effects which can be measured by numerical metrics.

Finally, in order for cyberinfrastructure to be adopted by scientists, it must be sustainable. There are good examples of projects for which funding is structured such that the project is sustained after the initial funds to build it are exhausted. These methods and structures should be codified and strengthened, and best practices communicated and followed. The need for sustainability exists both for “hard” cyberinfrastructure such as HPC facilities, networks, storage systems, and the like but also for the software tools which make the hardware useful and bind it together into a larger whole. It is difficult to build something long-lived in the fast-paced world of science and engineering. However, it is critical that scientists be able to rely on cyberinfrastructure both today and tomorrow, but also in the future when a new discovery changes the context of previously collected data, and additional insight can be gained from the reanalysis of the previous work in the field.

Consent Statement